Integrating Privacy in Architecture Design of Student Information System for Big Data Analytics

José Farnesio Huesca Barril, Qing Tan
School of Computing and Information Systems
Athabasca University, Canada
e-mail: jffhuesca@gmail.com, qingt@athabascau.ca

***Abstract*—Educational Data Mining (EDM) is an area of growing interest in academia with significant challenges and tremendous opportunities. Most EDM initiatives are based on finding patterns to aid and enhance student learning and performance, while others focus on program efficiency, service improvement, and college readiness. This paper is related to a case study being conducted at Sta. Teresa School, a high school in the Northern District of a West African country looking to improve service quality. By combining its Relational Database School Management System data sets with its anticipated online community forum, and aggregating it with Social media data, the school is expected to gain new actionable insights to enhance student services. In this paper, we present a model for incorporating privacy into big data analytics architecture integration with Social media, discuss some of the school's concern related to privacy and security, and offer some delivery options for its online community forum initiative.**

Keywords-educational data mining; big data analytics; data security and privacy; opinion mining; online community data mining; digital education, big data analytics archicture, cloud

I. INTRODUCTION

Schools in many African countries and other developing nations continue to face unprecedented policy and IT infrastructure challenges when undertaking digital transformation programs in education. This paper is related to a case study being conducted at Sta. Teresa School, a high school in the Northern District of a West African country, and addresses some of the privacy and security concerns faced by the school with its online presence and digital education initiatives.

Digital education is gaining popularity revolutionizing learning experience and dramatically changing student outcomes. A study conducted by Pwc Canada in 2014 reported that about 86 % of Canadian students would consider enrolling in a program that blends traditional and digital learning. The study also found that students' expectations for services to be delivered digitally will only increase in years to come [1]. These results appear to support a broader trend in education around the world; schools are increasingly investing in IT infrastructure and computing devices such as laptops, tablets, and other mobile technologies in favor of textbooks; shifting from the traditional system of teaching with blackboards and chalk, and paper-based assignments to the digital form of education [2].

Digital education is enabling the emergence of new student services to be offered and delivered online. These new services are creating new sources of data, offering new capabilities and opportunities for data collection and analysis that did not exist before [3]. These data sources are subject of study in Educational Data Mining (EDM).

EDM is the application of sophisticated data mining techniques to solving problems in education. One such technique is Opinion Mining or Sentiment Analysis- described as the computational study of people's opinions, sentiments, evaluations, attitudes, appraisal, views, emotions, subjectivity, etc., of a specific topic expressed in repositories to generate meaningful insights and discover knowledge [4] [5].

Responding to this new emerging trend in education, Sta. Teresa School has implemented a Relational Database (RDB) School Management System and integrated it with its online web services to provide a secure and easy mechanism for teachers, students and parents to access student records from anywhere, at any time. The system conceptual diagram is illustrated in Fig. 1.

Figure 1. School RDBMS online integration [6].

The school not only has received great feedback from the student community about its new online service, but also has observed a boost in its social media presence, and an increase in registration and profit margins.

The school is considering implementing an online community forum to capture and analyze student sentiments and feedback in order to enhance program curriculum and service delivery. It recognizes the added benefits of leveraging and augmenting social media data with its internal systems, but is concerned about: (1) the impact of the online

978-1-5090-4499-3/17/$31.00 ©2017 IEEE

system on its existing privacy and security policies, (2) how to comply with national and international obligations, particularly, data residency requirements and privacy laws, and (3) how to integrate the unstructured data from social media with its recently deployed RDBMS.

Integrating the school IT environment with social media networks requires careful consideration of the infrastructure architecture. Many Cloud Service Providers (CSPs) recognize these challenges and offer alternative hosted Leaning Management Solutions for schools. Considering its recent RDBMS investment, the School would also like to know which deployment option, on-prem, hybrid, or fully hosted cloud, will best address its three main concerns.

In what follows, we present a model for embedding privacy into big data analytics architecture design and integration with social media, then, we discuss deployment models (on-prem vs cloud) considering the school's existing RDBMS investment, and finally, we address the data privacy and security concerns, and suggest initiatives that the school can take to comply with and promote privacy laws.

II. ARCHITECTURE INTEGRATION

While most EDM tools deliver on the promise of privacy and security, they fail to integrate privacy into the architecture design. Many online services and data extraction tools rely heavily on network security transport protocols to provide data security, and business policies to safeguard individual's privacy. Twitter, for example, only allows SSL connections for Twitter API endpoints, it does this to protect user privacy as the information traverses across the public Internet [7]. In Jan 2016, Facebook moved from HTTP to an HTTPS based API (Graph API Webhooks), making it a requirement for all new callback Webhooks endpoints to use the new Graph API beginning April 2016 [8]. Graph API, the predecessor of Webhooks, was a low-level HTTP-based API that allowed developers to get data in and out of Facebook [9].

Although these mechanisms are important, they fail to protect individual's privacy when the network transport channel is tampered with or the slightest mishandling of business policy occurs, leading to the exposure of Personal Identifiable Information (PII). Accountable business practices, IT systems, and physical design and networked infrastructure should all integrate the *Privacy by Design* Principles to better address privacy issues [10].

A. Proposed Architecture

Our proposed architecture integrates privacy into the architecture design. We take a proactive approach to address PII data privacy during the data collection phase. Our solution for Sta. Teresa School integrates the school existing infrastructure with social media network through API calls

Given the equation

$$f(x) = get\ [Q(x),\ P(x)] \quad (1)$$

where

Q(x) => Data stream (i.e.: #stateresa, student records) and *P(x) => Strip off any PII (i.e.: remove [user_id, stud_id, screen_name])*

By implementing a privacy layer in the design, we can strip off any PII during data extraction. The proposed architecture shown in Fig. 2 implements various Big Data Analytics components such as:

- **Data Ingestion**-Batch, interactive, streaming, and real-time use cases
- **Data Management**-Store and process vast quantities of data
- **Data Integration**-Quickly and easily load data, and manage according to policy
- **Security**-Address requirements of Authentication, Authorization, Accounting and Data Protection.
- **Operations**-Organization wide operating procedures to enforce Security, Privacy and Compliance policies.

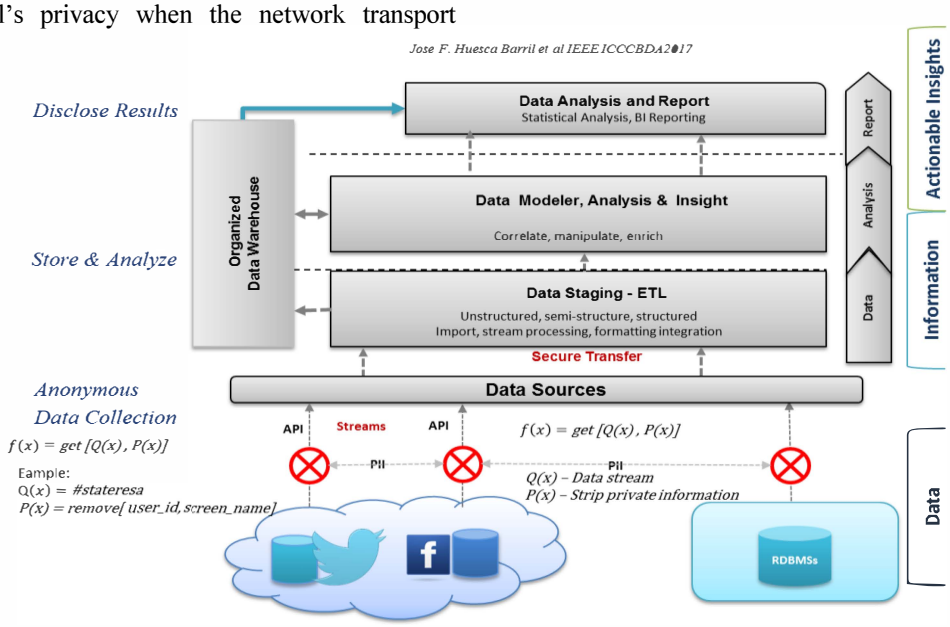


Figure 2. School online community social media analytics integration architecture.

Our proposed architecture implements a method for stripping off any PII data during data extraction to minimize any possible linking or identification. It also allows the school to stage the data sources before moving them into a data warehouse for analytics processing. This technique should take care of the school concern of privacy as well as unstructured and structured data integration. A very simple, but yet comprehensive approach of Text Mining using Twitter Streaming API and Python is provided in [11]. Additionally, an organization wide approach to data analytics should be adopted to guarantee the success of the analytics initiative and help the school mature in the DIKW (data-information-knowledge-wisdom) taxonomy.

B. Deployment Options

Integrating the school IT environment with social media networks requires careful consideration of the infrastructure architecture. Many CSPs recognize these challenges and offer alternative hosted Leaning Management Solutions for schools. Considering its recent RDBMS investment, the School would like to know which deployment option, on-prem, hybrid, or fully hosted cloud, will best address its three main concerns.

1) Fully hosted cloud-based solution

There are some fully hosted, cloud-based school management service providers that offer customers with a variety of attractive student services and data analytics projects, allowing customers to save on upfront cost in infrastructure, Fig. 3. Though, data analytics initiatives are usually not part of their standard offerings, and can be very costly. The Sta. Teresa School is mainly concerned about integrating its existing RDBMS school management system with unstructured data from its potential online community forum and social media streams and concerned about privacy and security implications.

Option 1: Fully hosted SaaS

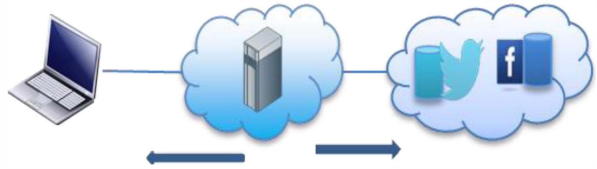
Figure 3. Fully hosted cloud-based solution.

With a fully hosted cloud service, the CSP has full control of the customer data and may not guarantee data residency requirements. Security and compliance responsibilities may also be provided by third parties that often are in compliance with privacy laws, but not necessarily safeguarding the customer's privacy.

Adopting this option would mean that the school would have to scrap its existing and successful School Management service and migrate to a completely new system. The cost on the user experience and to the organization might be too high and risky.

2) IaaS/PaaS and on- premise solution

Options 2 and 3 share some similarities, but also some differences. The main difference is primarily the hosting location of the new online community service: cloud (public/private), or on-prem. Option 2 is an on-prem solution and shall leverage the existing infrastructure environment, and connect through API to social media networks. With this option, the school retains and is in control of all the privacy and security policies, with the exception of the Social media data, however, once the data enters the school premises, the school policies can be enforced.



Option 2: On-Prem
Integrated with SMA

Option 3: IaaS/PaaS
Run from the cloud

Figure 4. Online community RDMS service.

Option 3, is a hybrid cloud approach. Similar to option 1, the online community development and deployment environment would be in the cloud, but unlike option 1, the customer remains in control of the service. With this option, the customer might decide to move its Student Management Service completely to the cloud (private) and remain in control of the service. However, it would still inherit most of the privacy and security restrictions of the cloud. It might also require an infrastructure project to link the school IT environment to the cloud.

III. PRIVACY CONSIDERATIONS

A. Privacy and Confidentiality

Privacy is about people. It is the control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others [12]. In the Big Data era, privacy concerns are not so much about who knows, but who should know [13]. Privacy can be so circumstantial that a person may not necessarily care if his or her contact information is seen on Yellow Pages, but overly concerned if the same information is seen on a police database. Similarly, one may have his or her date of birth (DOB) listed on Wikipedia or Facebook, but concerned if he or she receives an anonymous call inquiring about the same information.

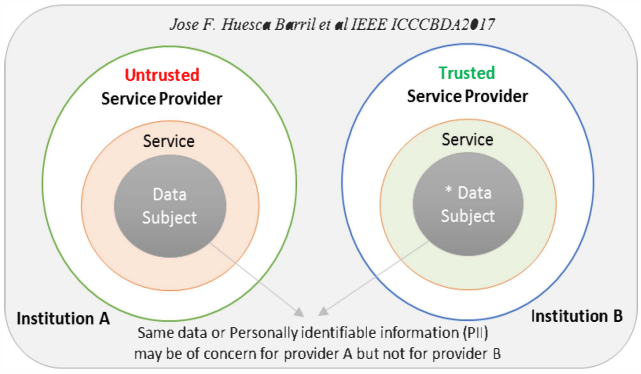


Figure 5. Data subject and service provider trust.

While privacy is about people, confidentiality is about data. In Fig. 5 we illustrate the concept of privacy and confidentiality, or the trust of whom to confide and or disclose data to. In the digital world, confidentiality refers to the property of a computer system whereby its information is disclosed only to authorized parties [14]. Though Wikipedia, Yellow Pages, and other organizations may not guarantee the integrity of one's data, one trusts and feels that they will not unlawfully or maliciously disclose or use the data, and will remain in compliance with the law.

B. *Privacy Preservation*

There are various privacy preserving data mining techniques and mechanisms to protect PII, such as: anonymization, de-identification, encryption, masking, randomization, pseudonymization and manipulation [15]. Each of these techniques serves a particular purpose. For instance, anonymization refers to manipulating, masking, or removing personal identifiers so that it is difficult or impossible to restore the original data; any remaining data cannot be linked to an individual. Encryption encodes the message or information using cryptographic algorithms so that only authorized parties can decrypt and read it. Often various techniques are combined to achieve a greater degree of privacy. As we argued above, most of these mechanisms fail to protect individual's privacy when the transport channel is tampered with or the slightest mishandling of policy occurs, leading to the exposure of PII. There are, however, situations where these mechanisms are needed, particularly, analysis that requires intervention to help individual students achieve better outcomes. Generally, a statistical analysis does not require the use of PII or reversible techniques for re-identifying individual subjects, which makes the collection of PII expendable.

C. *Privacy Protection and Security*

Students demand their data remain confidential, protected and in compliance, and the school must answer with adequate security measures for the data at rest and in transit. This means appropriate security control mechanism must be implemented for the collection, storage, use, and disclosure of the student data to prevent unauthorized users from accessing and tampering with it. A security or privacy breach could cost both the student and the school significant damage.

The damage could range from financial losses to social reputation. Fig. 6 illustrates common security threats and countermeasures that the School can take to help safeguard student data.

D. *Privacy and Compliance*

Compliance is the adherence to jurisdictional laws and regulations according to the context meaning, mainly location, date, and time in which the data is collected, used, and disclosed [17].

Sta. Teresa School does not currently face local pressure and tough requirements for student data privacy. The municipalities in which it presently operates are slowly adopting IT services, putting the school in a position to help shape local and state laws concerning student data privacy. Though, with an online presence, it is important that the school adopt, implement, and comply with data privacy and compliance laws.

In December 2013, the United Nations General Assembly adopted resolution 68/167 - The Right to Privacy in the Digital Age, which called upon all states to respect and protect the right to privacy in digital communication [16]. Enforcing this UN mandate is not often easy, particularly when the data is spread across different geographical and legal boundaries. In such scenarios, data ownership and usage purposes can be a challenge for policy makers; creating conflicting or overlapping policies, and making it even harder for researchers and industries to follow, comply and abide.

In 2014 alone, calls from privacy advocates in the US have led to nearly 100 bills being introduced in U.S. state legislatures to address issues of student privacy [18].

How must Sta. Teresa School approach online data privacy and compliance?

The Fair Information Practice, and the Organization for Economic Co-operation and Development (OECD) Privacy Guidelines are the most widely-accepted privacy principles, and also the foundation of privacy laws and related policies in many countries, (e.g., Sweden, Australia, Belgium) [19]. Additionally, with the help of local authorities, and state legislators, the school can follow a similar approach such as the one taken by the Canadian government to accommodate international data privacy laws.

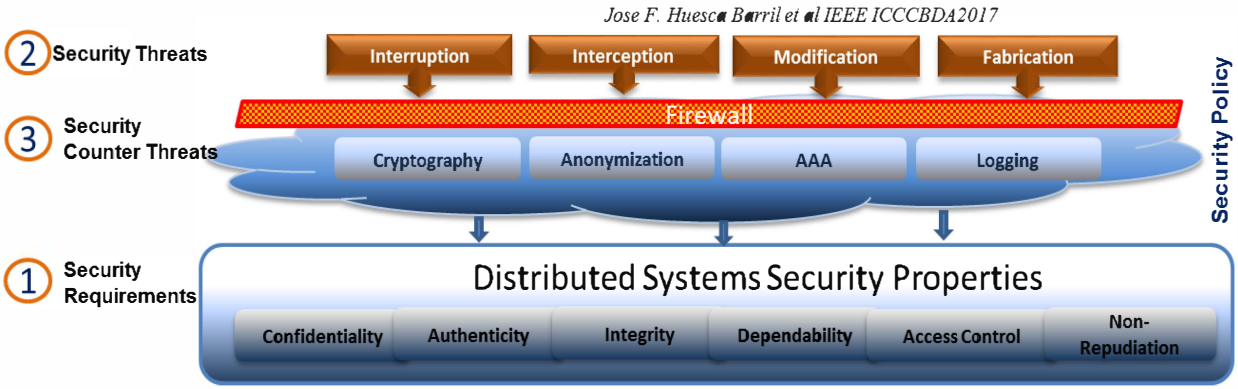


Figure 6. Security threats and countermeasures.

TABLE I. PRIVACY RECOMMENDATIONS FOR STA. TERESA SCHOOL

Point of concern	Proposed solution	Action items
Policy	Policy Activism	Work with local authorities, and state legislators to promote privacy laws.
Compliance	Education in Social Media [20]	Educate internal staff and students on the use of social media and the importance of compliance with national and international laws.
Confidentiality	Authorized access and proper use of data	Implement security measures to ensure data is maintained safely and strictly confidential.

In Canada, the Freedom of Information and Protection of Privacy Act (FIPPA), has been in effect in the province of Ontario since 1988 with the aim at protecting individual privacy while ensuring that records for public institutions are available to the public. The Canadian parliament also introduced a similar law in 2000, the Personal Information Protection and Electronic Documents Act (PIPEDA) to regulate how private sector organizations collect, use and disclose personal information in the course of commercial business

The PIPEDA act was also intended to reassure the European Union that the Canadian privacy law was adequate to protect the personal information of European citizens. As a result, EU countries and businesses are free to transfer personal information to organizations in Canada that are subject to PIPEDA or other Canadian laws, both at the federal and provincial level [21].

In the United States, under the American Health Insurance Portability and Accountability Act (HIPAA), PII rules apply to 18 specific identifiers shown in Table II below:

TABLE II. HIPAA PII [20]

Name	***Address***	***Birthdate***
Phone No	Fax No	Email Address
Social Security No.	Medical Record No.	Health Insurance Beneficiary No.
Account No	Certificate No.	Vehicle ID No.
Device ID No.	Personal URL	IP Address
Biometric ID	Facial Image	Other Unique ID Code

While there is no real pressure for Sta. Teresa School to comply with local or international data privacy laws, the influx of expat families taking long term assignments and temporary resident in the country can rapidly and significantly change its data privacy requirements. The School online presence and community forums are likely to expand its data domain into new dimensions, including open data and international student data, changing its data privacy requirements.

Some countries' privacy laws specify that the data be used solely for the purpose of which it was obtained [22]. However, this is not always the case. Often, privacy rights are cascaded from one organization to another, or even from individuals to individuals. In the past, for example, Facebook allowed individuals to volunteer their friends' status updates, check-ins, location, interests and more to third-party apps without the previous consent of the data owner [23].

In Table I above, we summarize some actions to be taken by the School around policy, compliance, and confidentiality.

IV. CONCLUSION

Educational Datamining is the application of sophisticated data mining techniques to solving problems in education. One such technique is Opinion Mining, which Sta. Teresa School, a high school in the Northern District of a West African country, hopes to use along with proper governance, architecture and deployment model, to improve its service delivery and increase its customer base. Because such data mining techniques often require the augmentation and/or aggregation of other data sources that may span different geographical boundaries and jurisdictions, privacy and security becomes major concerns. In this study, we have briefly exposed common security threat and counter measure mechanism to help the school address online threats, and proposed an architecture model that implements a privacy layer in the design.

While most EDM tools deliver on the promise of privacy and security, they fail to integrate privacy into the architecture design, and rely heavily on security transport protocols to provide data security, and business policies to safeguard individual's privacy. Although these mechanisms are important, they fail to protect individual's privacy when the transport channel is tampered with or the slightest mishandling of policy occurs, leading to the exposure of PII. By implementing a privacy layer in the design, we can strip off any PII during data extraction.

Further, we looked at three deployment options (on-prem, hybrid and fully hosted cloud) and contrasted them with the school privacy and security concerns taking into account its recent investment in a RDBMS solution, and recommended an architecture model to help it meet its technical challenges.

Finally, given the IT adoption rate in the area where Sta. Teresa School operates, we believe the School is in a unique position to help educate and influence privacy laws at both local and state level, and we have offered the School with some specific action items to help it enhance its privacy and security policies from an international and local perspective.

REFERENCES

[1] Pwc (2014, June 13). The connected classroom. How Canadians see the evolution of education. Retrieved Jan 15, 2017 from http://www.pwc.com/ca/en/public-sector-government/publications/pwc_citizencompass-theconnectedclassroom2014_june13.pdf

[2] Sabourin,J., Kosturko,L., FitzGerald,C., McQuiggan,S. (2015). Student Privacy and Educational Data Mining: Perspectives from Industry. In Proceedings of the 8th International Conference on Educational Data Mining (pp,164-170)

[3] Peña-Ayala, Alejandro. "Review: Educational Data Mining: A Survey And A Data Mining-Based Analysis Of Recent Works." Expert Systems With Applications 41.Part 1 (2014): 1432-1462. ScienceDirect. Web. 5 Dec. 2015

[4] Liu, B. (2012). Sentiment Analysis and Opinion Mining. New York: Morgan & Claypool Publishers

[5] Bouazizi, M., & Ohtsuki, T. (2015). Opinion Mining in Twitter. How to Make Use of Sarcasm to Enhance Sentiment Analysis. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

[6] School Automation System now a booming world. Retrieved on Nov 5, 2015 from http://claretalabs.blog.com/2012/03/school-automation-system-now-a-booming-world/

[7] Connecting to Twitter API using SSL. (n.d.). Retrieved Dec 2, 2015 from https://dev.twitter.com/overview/api/ssl

[8] Duppenthaler, M. (2015, October 30). Action required for new Graph API Webhooks (Real Time Updates). Retrieved December 2, 2015, from https://developers.facebook.com/blog/post/2015/10/30/real-time-updates-rebrand/

[9] Facebook. Graph API Overview. Retrieved Dec 02, 2015 from https://developers.facebook.com/docs/graph-api/overview

[10] Cavoukian A (2011). The Privacy by Design: 7 Foundational Principles. Retrieved April, 10, 2017 from https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[11] Adil Moujahid (2014). An Introduction to Text Mining using Twitter Streaming API and Python Retrieved. Dec 01, 2015 from http://adilmoujahid.com/posts/2014/07/twitter-analytics/

[12] Privacy and Confidentiality. Retrieved on Nov 28, 2015 from http://www.research.uci.edu/compliance/human-research-protections/researchers/privacy-and-confidentiality.html

[13] Qing Tan & Frederique Pivot (2015). Big Data Privacy: Changing Perception of Privacy. IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015. Pg 860-865

[14] Tanenbaum, Andrew S., Maarten Steen. Distributed Systems: Principles and Paradigms, 2nd Edition. Pearson Learning Solutions, 10/2006. VitalBook file.

[15] PII / PHI Data De-Identification. (n.d.). Retrieved December 3, 2015, from http://www.iri.com/solutions/data-masking/de-identification/overview

[16] GB979 Big Data Analytics Guidebook R14.5.1 - TM Forum. (2015). Retrieved December 10, 2015, from https://www.tmforum.org/resources/standard/gb979-big-data-analytics-r14-5-1/

[17] The Right to Privacy in the Digital Age. The Office of the United Nations High Commissioner for Human Rights. Retrieved Dec 03, 2015 from http://www.ohchr.org/EN/Issues/DigitalAge/Pages/DigitalAgeIndex.aspx

[18] S. Trainor, “Student data privacy is cloudy today, clearer tomorrow,” Phi Delta Kappan, vol. 96, no. 5, pp. 13–18, 2015

[19] Erika McCallister, E., Grance, Tim., & Scarfone, K.(2010). Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) (NIST Special Publication 800-122, 2010 Edition). Retrieved Dec 4, 2015, from National Institute of Standards and Technology Web site: http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

[20] Jon Dron. (2015). Embracing Social Media: A Practical Guide to Manage Risk and Leverage Opportunity [Review of Embracing Social Media: A Practical Guide to Manage Risk and Leverage Opportunity].Journal of Educational Technology & Society, 18(4), 525–528. Retrieved from http://www.jstor.org/stable/jeductechsoci.18.4.525

[21] PII / PHI Data De-Identification. (n.d.). Retrieved December 3, 2015, from http://www.iri.com/solutions/data-masking/de-identification/overview

[22] UK Information Commissioner's Office. Processing personal data for specified purposes (Principle 2). (n.d.). Retrieved December 13, 2015, from https://ico.org.uk/for-organisations/guide-to-data-protection/principle-2-purposes/

[23] Constine, J. (2015, April 28). Facebook Is Shutting Down Its API For Giving Your Friends’ Data To Apps. Retrieved December 7, 2015, from http://techcrunch.com/2015/04/28/facebook-api-shut-down/#.hcyvjxq:nael